提供：

databricks

データインテリジェンスプラットフォーム

for dummies
A Wiley Brand

インテリジェンスで
データと AI を民主化

AI で
エンタープライズデータ
を理解

ETL、DW、BI、AI で
イノベーションを加速

Ari Kaplan
Stephanie Diamond

Databricks 特集号

Databricks 社概要

Databricks 社は、データと AI の企業です。Comcast、Condé Nast、Grammarly、そして Fortune 500 の 60%以上を含む世界中の何千もの組織が、データ、アナリティクス、および AI を統合し民主化するために、Databricks のデータインテリジェンスプラットフォームを利用しています。サンフランシスコに本社を置き、世界中にオフィスを構える Databricks は、Lakehouse、Apache Spark ™、Delta Lake、MLflow のオリジナルクリエイターによって創立されました。詳細については、Databricks の各種ソーシャルメディアをぜひフォローしてください。


x.com/databricks


linkedin.com/company/databricks


facebook.com/databricksinc

データインテリジェンスプラットフォーム

Databricks 特集号

Ari Kaplan/
Stephanie Diamond 共著

for dummies®
A Wiley Brand

データ・インテリジェンス・プラットフォーム For Dummies®, Databricks 特集号

出版元:
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030 - 5774
www.wiley.com

弊社のその他の製品やサービスに関する基本情報、または読者の皆様の事業や組織向け「*For Dummies*」シリーズの作成につきましては、弊社米国事業開発部までお電話（877-409-4177）またはメール（info@dummies.biz）にてお問い合わせいただくか、www.wiley.com/go/custompubをご覧ください。製品またはサービス向けの「*For Dummies*」ブランドライセンスに関する情報は、BrandedRights&Licenses@Wiley.comまでお問い合わせください。

ISBN: 978-1-394-32363-0（ペーパーバック）; ISBN: 978-1-394-32364-7（電子書籍）; ISBN: 978-1-394-32365-4（ePub 形式）冊子版の一部の空白ページは、PDF 版電子書籍に含まれない場合があります。

謝辞

本書の出版にあたりご協力いただきました皆様に心より御礼申し上げます。

プロジェクトマネージャー兼エディター: Carrie Burchfield-Leighton

アキジションエディター: Traci Martin

上級クライアントアカウントマネージャー: Matt Cox

編集主幹: Rev Mengle

目次

はじめに

組織の成功は、データを効果的に活用し、インテリジェントな意思決定とビジネスの成長を促進することにより実現します。これには、分析や人工知能（AI）が利用可能な戦略的資産に生データを変換することが必要になります。AI を用いたデータインテリジェンスにより、スマートな意思決定を行うことができ、企業の成功につながります。

Databricks のデータ・インテリジェンス・プラットフォームは、データと AI のための統合型プラットフォームを提供し、組織によるデータの民主化と AI アプリケーションの構築を可能にします。チームが協力してデータのサイロ化を解消し、データドリブンな意思決定を行う文化が生まれます。従来型 AI や生成 AI、データウェアハウス、ビジネスインテリジェンス、ガバナンスの活用で、自社のデータ資産から最大限の価値を引き出してください。

本書の概要

データ・インテリジェンス・プラットフォーム *For Dummies*, Databricks 特集号では、企業の戦略をリアクティブなものからプロアクティブなものに変え、競争優位性を生み出す資産としてデータを活用する重要な手法についてご紹介します。本書では、以下の内容を取り上げます。

- データインテリジェンスの価値と AI の力
- Databricks データ・インテリジェンス・プラットフォーム
- 従来型 AI と生成 AI を併用したアプリケーション構築
- データ・インテリジェンス・プラットフォームが求められる理由

対象読者

執筆にあたり、本書の読者をいくつか想定しました。

- 複雑な問題の解決に AI を活用したいとお考えで、AI とデータを統合したソリューションをお求めの方。
- 意思決定を行う立場にあり、効率性、革新性、競争優位性を高める、統合型のオープンでスケーラブルなプラットフォームをお探しの方。

- データガバナンス、セキュリティ、規制コンプライアンスを確保する責任を担い、Databricks がどのように支援できるのか知りたい方。
- データ量や処理ニーズの増加に応じ効果的に拡張できるソリューションをお求めの方。
- 経営、戦略、またはミッションクリティカルな領域で、課題に取り組む新しい手法をお探しの方。
- Databricks のプラットフォームが既存のシステム、データインフラストラクチャ、分析ツールにどのように統合されているのか興味をお持ちの方。

上のいずれかに当てはまる場合、本書がお役に立てるはずです。

本書で使用するアイコン

本書では、重要性の高い情報に目を止めていただけるよう、アイコンを使用しています。各アイコンには、以下の意味があります。

「ヒント」のアイコンは、業務の簡素化、迅速化を図るための情報を見つけやすくするためのものです。

「ポイント」のアイコンは、記憶をたどる際に覚えておいていただきたい内容を示します。

「注意」のアイコンは、読者の皆様や会社に害を及ぼす恐れがある事柄にご注意いただくためのものです。

本書を読み終えた後で

本書は、データ・インテリジェンス・プラットフォームの知識を深めていただく上で有用ですが、より詳しい資料をお求めの場合は、以下をお勧めいたします。

- Databricks のデモ、製品ツアー、チュートリアルを視聴する。databricks.com/resources/demos よりご覧いただけます。
- 10 万名を超えるメンバーを擁する Databricks のコミュニティ community.databricks.com に参加する。

本章の内容

- データインテリジェンスの価値を知る
- データ・インテリジェンス・プラットフォームの主な機能について学ぶ
- 複数の業界におけるユースケースを見る

第1章 データインテリジェンスを理解する

データインテリジェンスを効果的に活用すれば、誰もがデータにアクセスしやすくなり、組織内の意思決定やデータ処理に革命をもたらすことが可能になります。データインテリジェンスに生成人工知能（AI）を組み合わせることで、データ分析を次のレベルに引き上げ、優れたインサイトの獲得と戦略的な意思決定が実現します。また、非技術系ユーザーでも、組織に関する質問を自然言語で尋ねることができるようになります。

データインテリジェンスとは、AI を応用して組織のデータの独自性を理解し、得られるインサイトの質を向上させ、実用的なインサイトを引き出す手法を指します。そのプロセスでは生成 AI を使い、膨大な量のデータを選別し、その意味を読み解き、意思決定に役立つインテリジェントなインサイトを導き出し、サービス、投資、事業戦略全般の改善を可能にします。

本章では、データを実用的なナレッジに変える力を組織に与えるデータインテリジェンスを定義し、それがもたらすメリットや影響について取り上げます。

データインテリジェンスとは

企業が競争優位性の獲得を目指すのであれば、自社のデータを理解しておかなければなりません。生成 AI により強化されたデータインテリジェンスが行う活動は、データの収集・分析から、データインサイトの適用による現実世界の問題の解決まで多岐にわたります。データインテリジェンスの活用により、企業は各種のパターンを見出し、トレンドを予測し、エビデンスを基に意思決定を下すことができます。このセクションでは、企業が成功するために必要なツールをデータインテリジェンスがどのように提供しているかを検証します。

インテリジェンス

データインテリジェンスでは、生成 AI とレイクハウスアーキテクチャを統合するメリットを組み合わせて、データの固有のセマンティクスを理解するデータインテリジェンスを強化します。これにより、Databricks のデータ・インテリジェンス・プラットフォームは、各ビジネスのニーズに合わせて、パフォーマンス最適化とインフラストラクチャ管理を自動化することができます。

シンプル

自然言語処理によって、ユーザーエクスペリエンスがシンプルになります。データインテリジェンスは、各組織で使われている言葉を理解するため、同僚に問いかけるような容易さで新たなデータの探索・発見が可能です。さらに、自然言語アシスタンス機能を通じて、コードの記述、エラーの修正、答えの発見を支援し、新たなデータやアプリケーションの開発を加速させます。

プライベート

生成 AI の利用拡大に伴い、データと AI の活用における、堅牢なガバナンスとセキュリティが欠かせません。Databricks は、ガバナンスとセキュリティを包括的にサポートするアプローチによって、エンドツーエンドの MLOps および AI 開発のソリューションを構築し、提供しています。データのプライバシーや知的財産の制御を犠牲にすることなく、OpenAI をはじめとする API の使用やカスタムモデルの構築などの AI イニシアチブを促進させます。

生成 AI とは、AI 自身が新しいコンテンツを解釈・作成できる人工知能の総称です。生成 AI のコンテンツには、テキスト、画像、動画、音楽、翻訳、要約、コードなどが含まれます。また、自由形式の質問に答えたり、チャットに参加したりするなど、特定のタスクを遂行できます。ChatGPT や DALL-E などのソリューションにより、一般の人々も生成 AI に触れるようになり、広く親しまれるようになりました。

ヒント

データインテリジェンスなくしてプラットフォームのスマート化は図れません。データウェアハウス（DWH）のようなプラットフォームは、高度なスキルを持ったエンジニアによる手動での保守と最適化を要するため、インテリジェントであるとは言えません。このような場合には、データインテリジェンスがプラットフォームの利用状況やトレンドを学習し、その学習内容をプラットフォームの改善と効率化に役立てることができます。

インサイトの収集、分析、応用

企業が競争力を高めるには、データの効果的な収集と解釈が不可欠です。データを効率よく収集して堅牢なデータセットに結合し、アナリティクスと AI でデータを分析、現実世界での意思決定を支援するのがデータインテリジェンスです。

企業によるデータ理解を支援

企業がデータを理解する上で鍵となるのはデータインテリジェンスです。高度なツールの活用で、顧客や市場に対する企業の理解が深まり、スマートな意思決定を下すことが可能になります。データインテリジェンスによりこれがどのように促進されるか、例をいくつか挙げます。

- **データ民主化の推進:** 自然言語を使うことで、テクノロジーに精通したプログラマーの助けを借りずとも、意思決定者自身が自社データについて質問できるようになったため、はるかに多くのユーザーが、企業のデータ資産から価値を引き出すことが可能になります。
- **業務の合理化:** データインテリジェンスは、各ビジネスのニーズに合わせて、パフォーマンス最適化とインフラストラクチャ管理を自動化します。
- **データガバナンスとコンプライアンスの確保:** データの理解とは、その出所を把握し、それをいかに活用するかを理解し、法的および倫理的な基準に準拠していることを確認することでもあります。データインテリジェンスにより、企業はデータガバナンスで効果をあげるツールを備えることができ、規制事項への準拠を担保するデータの品質とセキュリティを管理する上で役立ちます。

レイクハウスアーキテクチャをベースとして構築

データインテリジェンスで構築された統合型のオープンでスケーラブルなレイクハウスアーキテクチャが、データ関連の機能を単一でまとまりのある環境に統合する包括的なシステムとしての役割を果たします。

統合型プラットフォームの活用が、効率的なデータ管理と分析手法というメリットを企業にもたらし、データのサイロ化を解消し、あらゆるデータ資産を一元管理する単一のリポジトリを実現します。これにより、一貫性、正確性、ガバナンスを組織全体で確保できます。

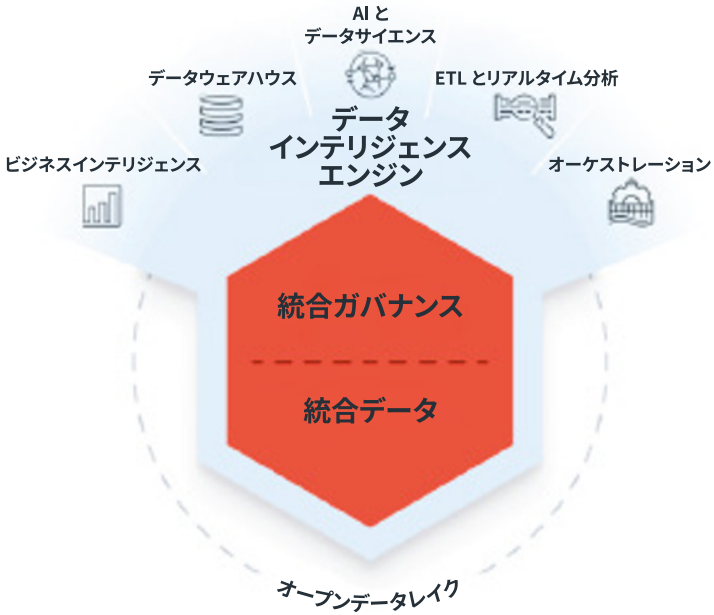
データインテリジェンスを最大限活用する

組織にとっての重要戦略として発展を遂げてきたデータインテリジェンスが、データの力を活用する上での支援となり、開発が簡素化されるため、人材に熟練度の高さを求めなくてよくなります。このセクションでは、こうした取り組みによりもたらされるメリットをまとめてお伝えします。

データの検索・理解が容易に

データインテリジェンスは、各組織で使われている言葉を理解するため、同僚に問いかけるような容易さで新たなデータの探索・発見が可能です。図 **1-1** に示すように、データインテリジェンスでは、単なるキーワードの一致にとどまらず、検索のコンテクストを理解することで、情報を容易に発見し、検索できるようになります。さらに、自然言語処理（NLP）ツールの採用で、ユーザーは平易な言葉でデータを照会できます。

*自然言語*は、データインテリジェンスにおいて極めて重要な機能であり、システムによる言葉の理解と解釈が可能になります。これにより、大量のテキストからセマンティクスを利用して重要な情報を簡単に抽出できるようになり、それを意思決定やカスタマーインサイトの向上に用いることができます。



出典: DATABRICKS

図 1-1: エンドツーエンドのデータプラットフォームのあらゆる段階で統一されたガバナンスとデータを追加する役割をデータインテリジェンスエンジンが担う。

サイロ化されたデータを単一のプラットフォームに統合

サイロ化されたデータを単一のプラットフォームに統合することで、多様なシステム、部署、場所にまたがるデータの断片化という、多くの組織が直面する問題に対処できます。データがサイロ化されると、それ以外の関連データから分離され、顧客の行動や市場動向といった概念の把握はほぼ不可能になってしまいます。

データの断片化は非効率性を生じさせ、さらに重要なことに管理職が全体像を把握できないために機会を逃してしまう可能性があります。データを単一の統合型プラットフォームにまとめておけば、このようなサイロが解消され、データが流れて分析可能になります。

統合型のデータプラットフォームでは、あらゆる構造化データ、非構造化データの保存と分析が可能な集中型リポジトリが構築されるため、データの正確性が担保され、利用する高度な分析や AI アプリケーションで高い効果が得られます。それが、最大限のデータ資産活用という結果に現れるのです。

非技術系ユーザーがデータからインサイトを得られるよう支援

データへのアクセス簡素化とは、非技術系ユーザーでもデータへのアクセスを容易にするということです。データへのアクセスが容易になれば、IT 部門に頼ることなくインサイトを得ることができます。すべての従業員がデータ分析の基礎と、それを容易にする利用可能なツールを把握していることを確認してください。

企業運営の効率化とコスト削減の推進

データインテリジェンスを活用することで、企業のテクノロジー業務が合理化され、コスト削減につながります。また、予測分析によるトレンド予測と企業戦略の調整が可能になります。


ヒント

AI により、時間のかかる手動プロセスの自動化など、テクノロジーの効率化とコスト削減の新たな機会を見出すことができます。例えば、不適切に利用されているリソースを特定して再配分したり、より大規模かつ高速に拡張できるようデータの保存方法を改善したり、などが挙げられます。重要なのは、生データのソフトウェアエンジニアリングから分析に至るまで、テクノロジーのあらゆる段階で、AI による業務改善が可能になったということです。

コラボレーションの促進

タスクすべてに共通の環境を用意することで、さまざまな部門間の連携を促進するのが各種のデータインテリジェンスツールです。部門が違っても、同じデータセットで同時に作業を行い、コードを共同開発し、ダッシュボードやレポートからのインサイトを共有して、一体的に意思決定を下すことができます。この協調的な環境が、共通の目標へと向かう取り組みを促します。

ビジネス全体に与える影響

データインテリジェンスは、ビジネスのあらゆる面で機能性と効率性を高めます。機能性と効率性の進展を促し、人のニーズや倫理基準への確実な対応が図られるため、データと AI のエコシステム全体が強化され、複雑な課題にも対処できるようになります。このセクションでは、データインテリジェンスがこのような状況を作り出す主な方法について説明します。

データの品質と完全性の向上

AI システムと分析プロセスのいずれでも、その有効性の基礎になるのが、データの完全性と品質です。異なるデータソース間でのデータ検証、クレンジング、一貫した管理のためのメカニズムをデータインテリジェンスが担うことで、こうした面が強化されます。

イノベーションと新たなビジネスモデルの推進

イノベーションの実現と新たなビジネスモデルの構築で極めて重要になるのが、データインテリジェンスです。データの分析により、企業は新たなトレンドや未開拓の市場ニーズを特定でき、革新的な製品やサービスを新たに生み出す機会が生まれます。


ヒント

データドリブンなアプローチにより、サブスクリプションサービスやオンデマンドプラットフォームなどのビジネスモデルを試すことが可能になり、競争上の優位性を得ることができます。データから得られるインサイトが、新たな収益源や革新的な戦略につながることもあるでしょう。

AI と ML の導入を加速

データインテリジェンスが、データを AI と機械学習（ML）が利用できる形式に整えて変換し、AI と ML の基盤を提供します。正確かつ信頼性の高い AI モデルをトレーニングするには、高品質で適切に管理されたデータが不可欠です。

データ・インテリジェンス・プラットフォームの主要機能を評価する

データ・インテリジェンス・プラットフォームは、企業がデータを価値あるビジネス資産とする上で必要になるツールを提供する役目を果たします。これらのツールがデータを統合プラットフォームに集めて分析し、効果的な戦略を策定します。こうしたプラットフォームで何ができるのかを押さえておくと、データのニーズと目標に適したプラットフォームの選定に役立ちます。


ヒント

データ・インテリジェンス・プラットフォームを評価する際、組織は拡張性、パフォーマンス、使いやすさ、統合機能などの要素を考慮し、選択したプラットフォームが特定のビジネス要件や技術インフラに適合していることを確認する必要があります。

NLP の活用

翻訳ソフトウェア、チャットボット、検索エンジンといったツールの中核をなすテクノロジーが、NLP です。NLP を活用することで、顧客レビュー、ソーシャルメディアの投稿内容、サポートチケットなど、企業に存在する非構造化データが持つ潜在能力を、データ・インテリジェンス・プラットフォームが最大限に引き出せるようになります。

データセキュリティと拡張性を成長に合わせて確保

セキュリティと拡張性は、あらゆる組織の成長にとって不可欠です。データインテリジェンスは、個人情報の保護、データ関連の規制に対するコンプライアンスを確保する強力なセキュリティ機能を備えています。また、増え続けるデータと組織の拡大するニーズに対応できるよう、拡張性も備えていなければなりません。

多様なスキルレベルに対応可能なプラットフォームの構築

データ・インテリジェンス・プラットフォームは、さまざまなレベルの技術的専門知識を持つユーザーが利用できるものでなければなりません。これをクリアしていれば、データサイエンティストからビジネスアナリストまで、組織の幅広いユーザーによるプラットフォームの活用が可能になります。

ノーコードツールと直観的なインターフェースによるユーザーエクスペリエンスの簡素化で、データの利用範囲が広がり、より多くのステークホルダーが、情報に基づくデータドリブンな意思決定を行うことができるようになります。

データプロセスの自動化

データ・インテリジェンス・プラットフォームの自動化機能は、企業が膨大なデータを処理する方法を根本から変えます。データプロセスに自動化を組み込むことで、効率性、精度、スピードは大幅に向上します。自動化によりワークフローが合理化され、手作業が減り、データ管理が全体的に向上します。

手動によるデータ処理の必要性が減るというのは、自動化のメリットとして極めて重要なもののひとつです。手作業は時間がかかる上、エラーが発生しやすいものでもあります。自動化で人による入力の必要が最小限になり、データ処理上のミスが起こりにくくなります。

データの収集、クレンジング、処理といった作業が自動化されれば、業務効率が改善し、時間の節約になるのに加え、戦略的な取り組みにリソースを割くことができます。

さまざまな業界におけるデータインテリジェンスのユースケースの考察

データインテリジェンスは、金融、ヘルスケア、エネルギーなど、さまざまな業界で活用され、データドリブンなインサイトの活用が、ビジネスのあり方を変革しています。このセクションでは、データインテリジェンスが企業の顧客理解、プロセスの改善、不正行為の検出などにどのように役立っているかを示す例をいくつかご紹介します。

- **金融:** 金融リスクの管理、経済動向の予測、規制の遵守にデータインテリジェンスを活用しています。銀行をはじめとする金融機関は、データを分析して信用力を評価し、不正行為を特定し、顧客を分類します。
- 小売・消費財: データインテリジェンスを活用して顧客の嗜好を理解し、在庫管理を改善し、サプライチェーンを最適化し、個々の顧客に合わせたマーケティングを行います。
- **政府・公共機関:** 公共機関では、データインテリジェンスがサービスの向上と情報に基づいた政策決定に不可欠です。行政機関は、経済状況の変化を監視し、サービス提供を改善するためにデータを使用しています。
- **保険:** 保険会社では、リスクの評価、保険料の設定、不正請求の検出にデータインテリジェンスを活用しています。大量のデータを学習することで、リスクをより明確に把握し、保険金請求プロセスを効率化します。
- **ヘルスケア:** 医療機関では、データインテリジェンスを活用して、患者ケアの強化、コスト管理、研究を行っています。データ分析は、医療上の意思決定をサポートし、効果的な治療法の特定に役立ちます。
- **エネルギー:** エネルギー分野では、企業はデータ分析を使ってエネルギー使用量を監視・予測し、送電網の効率を向上させています。

データインテリジェンスのアプリケーションは業界によって異なるかもしれませんが、共通の目標は、データから価値あるインサイトを引き出し、それを活用してビジネスの成長や顧客体験の向上を図ることです。

本章の内容

- レイクハウスがない場合に生じる課題の考察
- レイクハウス、データウェアハウス、データレイクについて知る
- 生成 AI と従来型 AI の考察
- AI の持つ可能性を押さえる
- レイクハウスアーキテクチャと生成 AI の活用
- データ・インテリジェンス・プラットフォームを導入する

第 2 章
レイクハウス、生成 AI、従来型 AI について知る

生成人工知能（AI）を備えるレイクハウスアーキテクチャ上に構築されたデータ・インテリジェンス・プラットフォームは、組織全体でデータと AI の民主化を推進する強力な手段となります。レイクハウスアーキテクチャは、膨大な量の構造化、非構造化データをまとめてひとつの統合環境に保存・処理でき、データウェアハウス、ビジネスインテリジェンス、従来型 AI、生成 AI を新たなレベルへと引き上げます。

本章では、レイクハウス、データウェアハウス（DWH）、データレイクの違い、そしてレイクハウスに生成 AI と従来型 AI が加われば、組織にとってその価値がどのように高まるのかを検討します。

レイクハウスがない場合に生じる課題

企業のほとんどが、ビジネス目標を達成する上で、データと AI を効果的に組み合わせるのは難しいことだと感じています。こうした課題には、データインテリジェンスのエコシステムに欠かせないさまざまな要素が関わってきます。データ管理と AI の統合で直面する問題には、以下のようなものがあります。

- データと AI のサイロ化。データがサイロ化すると、運用コストの高騰を招きます。
- データのプライバシーと管理上で生じる問題。一貫性に欠けるポリシーは、データに対する信頼性を低下させます。
- 技術的専門性の高い人員への依存。異種のツールが、チーム間で共同作業をする際の生産性を低下させます。

物事を機能させるには、複数のサービスをつなぎ合わせる必要があり、要素それぞれに克服すべき課題があります。図 2-1 は、左上のデータレイクから時計回りに各要素とその課題を示したものです。

- **データレイク:** データレイクの課題は、膨大な非構造化データの保存と管理にあります。
- **機械学習（ML）:** 複雑なアルゴリズムの精度を確かめ、開発、適用、監視することが課題になります。
- **ストリーミング:** 連続するデータストリームをリアルタイムで処理するという、高い技術的要求があります。
- **生成 AI:** AI 技術で新しいリアルなコンテンツを生成するには、複雑さという課題があります。
- **データウェアハウス:** 問題は、分析用に構造化データを一元管理することですが、複雑でコストがかさみかねません。
- **ビジネスインテリジェンス（BI）:** 難しいのは、データを効果的に視覚化し、ビジネス上のインサイトを抽出できるかどうかという点にあります。
- **オーケストレーションと抽出・変換・ロード（ETL）:** これには、データの準備と移動の調整が必要になります。

- **ガバナンス：**各種規制を遵守しながら、強力なデータ管理とセキュリティ対策を実施することが課題になります。
- **データサイエンス:** 科学的手法を用いてデータを探索・分析する際のタスクは複雑なものになります。

図 2-1: データインテリジェンスエコシステムの課題。

これらの要素の詳細と、その課題に Databricks のデータ・インテリジェンス・プラットフォームでどのように対処するかについては、第 3 章をご覧ください。

データ・インテリジェンス・プラットフォームには、あらゆるデータタイプの保存・管理を行うオープンなデータレイク、データの信頼性を確保し共有を可能にする統合データストレージ、統合型のセキュリティ、ガバナンス、カタログ環境、データのセマンティクスを理解する AI 搭載エンジンなど、多くの要素で組織全体を支援し、データサイエンス、AI、ETL、リアルタイム分析、オーケストレーション、データウェアハウスのエクスペリエンスを向上させます。

レイクハウスとデータウェアハウス/データレイクの比較

レイクハウスでは、データのストレージと分析に、DWH やデータレイクとは異なるアプローチが採用されています。DWH やデータレイクからレイクハウスへの移行を促したのは、膨大な量の構造化データと非構造化データをよりスケーラブルでオープン、かつコスト効率に優れたソリューションで管理する必要性です。このセクションでは、これらの違いを見ていきます。

オープンアーキテクチャ

Databricks では、独自の形式や閉鎖的なエコシステムに縛られることなく、データを常に管理下に置くことができます。Databricks レイクハウスは、広く採用されているオープンソースプロジェクト Apache Spark、Delta Lake、MLflow によって支えられています。さらに、Delta Sharing は、レイクハウスからのライブデータをセキュアに共有するためのオープンなソリューションです。あらゆるコンピューティングプラットフォームでの共有を可能にします。コストのかかるレプリケーションや複雑な ETL は必要ありません。

統合型アーキテクチャ

データの統合、ストレージ、処理、ガバナンス、共有、分析、AI のためのレイクハウスアーキテクチャ。構造化データと非構造化データを扱うための単一アプローチ。データリネージとデータプロビナンスのためのエンドツーエンドの単一ビュー。すべての Python、R、Scala、SQL に対応する単一ノートブック。バッチとストリーミングに対応する単一ソース、主要な 3 大クラウドプロバイダすべてに対応する単一プラットフォーム。

スケーラビリティ

レイクハウスは、従来型の DWH やデータレイクと比べて、拡張性が高く、低コストで高いパフォーマンスを実現しながら、数兆ものレコードまで拡張可能です。パフォーマンスとストレージの自動最適化により、世界最高水準のパフォーマンスを提供し、あらゆるデータプラットフォームのなかで最も低い総所有コスト（TCO）を実現します。

データガバナンスとセキュリティの改善

レイクハウスは、組織全体のすべてのデータと AI アクセスに対し単一のセキュリティおよびガバナンスモデルを使用することで、データの管理と保護を行う手法を改善します。従来の DWH やデータレイクでは、複数のガバナンスソリューションが断片化し、さまざまな種類のデータが存在するため、一貫したポリシーや保護の適用が困難ですが、統合型のガバナンスプラットフォームがひとつあれば、規制の遵守が容易になり、データをよりセキュアに保つことができます。

生成 AI と従来型 AI の違い

AI には、生成 AI と従来型 AI の 2 種類があります。従来型 AI は、各店舗の今後の売上予測や、数百万もの顧客を異なるセグメントにグループ化するなど、数的予測やアイテムの分類が必要な分野で有用です。一方、生成 AI は、PowerPoint、PDF、Word 文書などの非構造化データからテキストで要約を作成する、一般的な質問に回答するといった場合に有用です。

生成 AI は、トレーニングデータから学習したパターンを基に新たなコンテンツを生成し、データの単なる分析にとどまらず、テキスト、画像、音声、動画、その他のメディアを解釈して検索を行うことが可能です。また、社内で行うテキストの新規作成やソフトウェアコードの記述・編集にも利用できます。

生成 AI で鍵を握るのが、スタイルと構造という点で、トレーニングデータに類似しながら、直接のコピーではない新たなコンテンツを生成する機能です。大規模言語モデル（LLM）などの生成モデルは、データに内在するパターンを特定し、そこから学習することで、概念を組み合わせて独自に内容を創り出すことができます。ビジネスでの生成 AI 活用例には、以下のようなものがあります。

- **テキストの生成:** 人間のような文章や説明文を書くことができ、訓練には企業の独自データを使います
- **テキストの要約:** 大量の文書を取り込み、人間が容易に解釈できる大意や評価を提示します
- **ソフトウェアコードの記述:** シンプルなプロンプトに従い、SQL、Python、Scala、R などのコードを記述します
- **データ資産の文書化:** テーブルやカラムの内容を記述して、セマンティック検索を向上させます

データインテリジェンスの強化における AI の重要性を認識する

AI とデータインテリジェンスの組み合わせは、企業がデータを分析、理解、活用する手法を大きく進歩させます。この効果により、組織が戦略を立てる能力が向上し、市場の変化や消費者のニーズへの迅速な対応が可能になります。

データインテリジェンスの改善における AI の重要性は、その技術面にとどまらず、さまざまな分野にわたりイノベーションを推進する能力にあります。

レイクハウスアーキテクチャと生成 AI の活用

レイクハウスアーキテクチャと生成 AI を統合すると、双方のテクノロジーが持つ能力が強化され、データ処理と分析をさらに強力な環境で実施できます。

レイクハウスアーキテクチャを活用する

レイクハウスアーキテクチャは、データレイクとデータウェアハウスそれぞれの優れた機能を組み合わせることで、データストレージと管理基盤を実現します。これにより、構造化データと非構造化データの保存先が単一のリポジトリであっても、組織は分析や ML のタスクを実行することができます。

生成 AI を組み込むことで、データドリブンなインサイトを分析・生成する新たな手法が得られます。このテクノロジーが、組織によるデータ品質の改善、より正確な予測モデルの開発を支援し、競争上の優位性につながります。

レイクハウスアーキテクチャと生成 AI の統合により、組織はデータ管理の有効性を高め、データドリブンな意思決定の新たな可能性を見出すことが可能になります。

オープンデータストレージを利用する

オープンデータストレージを利用することで、信頼性とデータ共有の利便性が向上します。レイクハウスアーキテクチャを採用し、生成 AI の機能を導入する上で、オープンデータストレージは不可欠で、組織による生成 AI アプリケーションの容易かつ効率的な開発・デプロイを支援するデータストレージレイヤーを実現するテクノロジーです。その活用が、生成 AI の力を最大限に駆使し、イノベーションを推進することを可能にします。

レイクハウスに生成 AI の機能を統合する

生成 AI の機能をレイクハウスアーキテクチャに統合すると、データ分析が強化されます。（データレイクとデータウェアハウスを組み合わせた）レイクハウスの強みを活用することで、組織はデータインテリジェンスの取り組みを推進できます。この統合には、主に以下のようなメリットがあります。

- **データタスクの自動化:** レイクハウス内のデータ操作を生成 AI が合理化します。従来型 AI には、データクレンジングの自動化が可能ではありますが、生成 AI ではさらに、モデルのテストやトレーニング用の人工データ生成による支援で、分析と AI アプリケーションの堅牢性を担保できます。
- **検索機能の強化:** AI により、レイクハウス内でのインテリジェント検索機能が強化され、ユーザーが、自然言語クエリを活用して、データ資産間の関係を効率よく見出し把握できます。これにより、単なるキーワード検索を超えたデータの発見が容易になり、分析に適したデータセットへ簡単にアクセス可能になります。
- **カスタム AI アプリケーションの開発:** AI をレイクハウスのフレームワークに統合すると、特定のニーズに応じたアプリケーションを創り出すことができます。例としては、自社データを基にした LLM の構築、予測モデルの開発、レコメンドエンジンのカスタマイズ、複雑なレポート作成作業の自動化などが挙げられます。


ポイント

生成 AI の機能をレイクハウスのアーキテクチャに統合することで、組織はデータからより多くの価値を引き出すことが可能になります。

データチームのコラボレーションを実現する

レイクハウスのアーキテクチャと生成 AI の機能が融合すると、データチームによるコラボレーションの可能性が大きく高まり、意見交換が盛んになるため、イノベーションの文化が育まれます。データチームは協力して、AI モデルの構築、トレーニング、デプロイを効率よく進めることができ、レイクハウスによるデータ管理と生成 AI の創造的な可能性という強みが、ビジネスの成長を促進する上で活かされます。

AI でデータ分析とインサイトを強化

AI によるデータ分析の強化を図るには、AI を搭載した多様なツールや手法を駆使して、データ分析のプロセスにある複数のステージを自動化・効率化します。データ分析のさまざまな段階に、以下のような方法で AI を統合可能です。

- **データ準備:** AI は、データのクレンジング、整理、前処理といったデータの準備段階を自動化できます。AI ツールは、データ品質の問題を検出して修正する、非構造化データから情報を抽出する、形式が異なるデータを結合するといったことが可能です。
- **データ探索:** AI アルゴリズムは、自然言語を使いデータを探索できるため、人間には一見分かりづらいようなインサイトでも見つけ出すことができます。
- **データ解釈:** AI は、データから要約、インサイト、ストーリーを生成して、データのより深い解釈を支援します。また、データを基に、因果関係を特定し、将来の結果や行動を予測することも可能です。
- **データ品質:** AI は、データやモデルの品質にある偏りを検知し、自動でフラグを立て、修正を支援することができます。

データ関連の複雑なタスクとプロセスを自動化

データ関連の複雑なタスクとプロセスの自動化とは、作業にテクノロジーを活用し、データの処理と分析の効率化を図るという意味です。この自動化により、データレイクにある大量の非構造化データが整理され、生成 AI と ML モデルが構築・適用され、データフローが途切れなくリアルタイムで処理されるようになります。

ジョブのオーケストレーション（複数の IT 自動化タスクやプロセスを調整・実行すること）では、要件を満たす適切なインスタンスと開始時間を AI が自動で選択し、オートスケーリングやエラー修正といったタスクを処理してくれます。

テーブルのファイルサイズ最適化など、データエンジニアリングでは AI からメリットが得られる領域が多く、自動化することも可能です。データの読み書きに最適なファイルサイズの割り出しに、エンジニアは多大な時間と専門知識を費やすものですが、それは、パフォーマンスの大幅な改善が期待できるからです。この複雑なタスクを自動化できれば、革新と言ってもよいでしょう。

ETL 処理でインテリジェントなオートスケーリングを行うと、クラスターの使用率が最適化され、ストリーミングワークロードで生じるエンドツーエンドのレイテンシー（データ転送の要求が処理されるまでの時間的遅延）を最小に抑えることができますが、これは、データの量と処理上のニーズを基に、指定した上限までリソースを自動調整することで実現します。処理速度が追いつかずデータの到着に時間がかかる場合には、スケールアップを効率よく行い、負荷が低い場合であれば、スケールダウンして、シャットダウンの前に確実にタスクを完了させるため、インフラストラクチャにかかるコストを削減できます。

データ・インテリジェンス・プラットフォームの導入

データサイエンティスト、データエンジニア、アーキテクト、ビジネスアナリストなど、さまざまな役割を担うユーザーが活用可能な統合型プラットフォームを実現するデータ・インテリジェンス・プラットフォームは、組織のイノベーションを推進する力となり、データのさまざまなステージを単一の統合環境に結合します。以下によって、この統合が可能になります。

- **データ統合:** さまざまなソースから、データを 1 か所にまとめることができます。データベース、データウェアハウス、データレイク、ストリーミングデータソースからのデータ統合に対応しているため、あらゆるデータを単一のプラットフォームで容易に操作できます。
- **処理・分析:** データがプラットフォームに取り込まれれば、処理と分析が可能になります。このプラットフォームは、Python、R、Scala、SQL など、使われると考えられる主要な言語をすべてサポートしています。搭載された機能やライブラリで、データのクレンジング、変換、分析を容易に実施できます。
- **ワークスペースでの共同作業:** さまざまなチームメンバーが同じデータやプロジェクトで共同作業できる共有ワークスペースをプラットフォームが提供します。データエンジニア、データサイエンティスト、アナリストが、ビジュアライゼーションやダッシュボードをすべて同じプラットフォーム上で作成できます。ワークスペースを共有することで、バージョン管理が容易になり、最新のデータと分析機能を誰もが利用できます。

- **作業する場所の統一性:** データへのアクセス制御、リソース管理、ジョブの監視など、データ分析のワークフロー全体を単一の場所から管理できます。リソースが適切に配分され、進行中のジョブ監視が容易になります。
- **シームレスなデプロイ:** 構築したデータ分析ソリューションは、本番環境へ容易にデプロイできます。シームレスなデプロイが可能なため、データ・インテリジェンス・プラットフォームを利用しない場合に発生しがちな問題に悩まされることなく、開発から本番環境へデータプロジェクトを移行できます。

このプラットフォームについては、第 3 章で詳しく取り上げます。

本章の内容

- » Databricks のデータ・インテリジェンス・プラットフォームをレビューする
- » Databricks プラットフォームアーキテクチャの考察
- » 開発者の取り組み支援

第 3 章
Databricks データ・インテリジェンス・プラットフォームの使用を開始する

企業は常に、データアーキテクチャを簡素化しながら、有意義なインサイトを獲得する能力をどのように向上すればよいかを模索しています。本章では、Databricks のデータ・インテリジェンス・プラットフォームを導入いただく際に、押さえておきたい基本事項を取り上げます。

Databricks データ・インテリジェンス・プラットフォームのご紹介

Databricks のデータ・インテリジェンス・プラットフォームは、組織全体でのデータと人工知能（AI）の活用を促進します。レイクハウスを基盤とするプラットフォームが、あらゆるデータ、AI、ガバナンスの要件をサポートするオープンな統合環境を提供し、インテリジェンスエンジンが、データの特性を理解します。

Databricks は、抽出、変換、ロード（ETL）やデータウェアハウジング、生成 AI に至るまで、データと AI のジャーニーをシンプルにし、目標達成の迅速化を可能にします。

DatabricksIQ によるデータインテリジェンスの実現

Databricks では、生成 AI のパワーとレイクハウスアーキテクチャの包括的な機能を併せ持つ、DatabricksIQ と呼ばれるデータインテリジェンスエンジンを開発します。DatabricksIQ が、お客様のビジネス、データに特有のニュアンスを学習し、幅広いユースケースに対応した自然言語によるアクセスを実現し、組織のすべての従業員が、自然言語でデータを検索、理解、照会できます。DatabricksIQ は、データ、利用パターン、傾向といった情報から、お客様が業務で使用する用語と独自のデータ環境を把握します。生成 AI の一種であり、翻訳、要約、質問への回答、テキスト生成など、幅広い言語関連のタスクを実行できる大規模言語モデル（LLM）を単純に利用するよりはるかに優れた回答が得られます。

当然、LLM が言語というインターフェースでデータの扱いを可能にするため、多数のデータ関連企業が AI アシスタントを採用しつつありますが、実際のところ、こうしたソリューションの多くは、エンタープライズデータを処理する上で不十分です。どんな企業にも、業務関連の質問に答えるのに必要な独自のデータセット、専門用語、社内ナレッジがあるため、インターネットで訓練を受けただけの LLM で答えを得ようとすると、結果に誤りが生じることが珍しくありません。顧客の定義や会計年度といった単純なことでも、企業によって異なります。

DatabricksIQ は、企業全体の業務とデータの概念を自動で学習することで、この問題を直接解決するデータインテリジェンスエンジンです。Databricks プラットフォーム全体のシグナル（Unity Catalog（UC）、ダッシュボード、ノートブック、データパイプライン、ドキュメントなど）を利用して、プラットフォームが独自に備えるエンドツーエンドの性質を活用し、実際にどのようにデータが使用されているか確認し、高度に専門化された正確な生成 AI モデルを構築します。

自然言語を介したシンプルなユーザーエクスペリエンス

Databricks では、自然言語処理（NLP）を活用し、ユーザーの皆様の利便性を大幅に向上しています。組織で使われる専門用語を理解するよう設計された Databricks のデータ・インテリジェンス・プラットフォームにより、同僚に問いかけるような容易さでデータの検索や発見を行うことができます。

NLP は、システムによる自然言語の理解と解釈を可能にします。この機能は、新たなデータアプリケーションの開発にも応用可能です。コードの記述、エラーの修正、質問への回答を支援することで、開発プロセスがスピードアップします。

プライバシーとガバナンスの確保

これまでにないほど重要性が高まっているのが、データと AI アプリケーションに求められる強力なガバナンスとセキュリティです。ガバナンスとセキュリティに対し、Databricks では、統一されたアプローチで機械学習運用（MLOps、機械学習モデルをビジネスに適用する上で必要な開発、分析、運用を効率化するための手法）と AI 開発をサポートし、包括的なソリューションをお届けします。このソリューションにより、幅広い分野で AI の取り組みを推進でき、プライバシーを確保し、知的財産を管理いただけます。

MLOps は、ML エンジニアリング機能のひとつで、ML モデルを本番環境に移すプロセスの合理化と、データの変更に応じてモデルを維持・監視することに重点が置かれます。

データ・インテリジェンス・プラットフォームの活用

Databricks が開発した、データレイクハウスと生成 AI の力を活用するデータ・インテリジェンス・プラットフォームでは、AI が持つ可能性をデータレイクハウスプラットフォームの中で探求する機能が大きく進歩しました。Databricks のデータ・インテリジェンス・プラットフォームが他に比べ際立つのは、統合されたガバナンスレイヤーがデータと AI の両方をカバーしている点にあります。また、ETL、SQL、ML、ビジネスインテリジェンス（BI）すべてに単一のクエリエンジンで対応できます。さらに、Mosaic AI の統合により、DatabricksIQ をサポートする AI モデルの開発も可能です。この統合は、組織の誰もがデータへアクセスできるようにする上で不可欠なものです。

レイクハウスという概念を生み出した Databricks が、あらゆるデータとガバナンスのニーズに対応するオープンで統一されたアーキテクチャをお届けし、単一システムでの構造化データと非構造化データの保存・管理までをも実現します。

Databricks では、低コストながら非常に高速なクエリ性能を発揮する次世代エンジン Photon の開発といったパフォーマンス強化にも注力し、プラットフォームの拡張性と効率性の向上を図り、最も負荷の高いデータ処理にも対応が可能です。

図 3-1 は、Databricks データ・インテリジェンス・プラットフォームのアーキテクチャを分かりやすく示したものです。このセクションでは、各要素を図の下にあるものから順に、どのように組み合わさってプラットフォームを形作っているかを見ていきます。

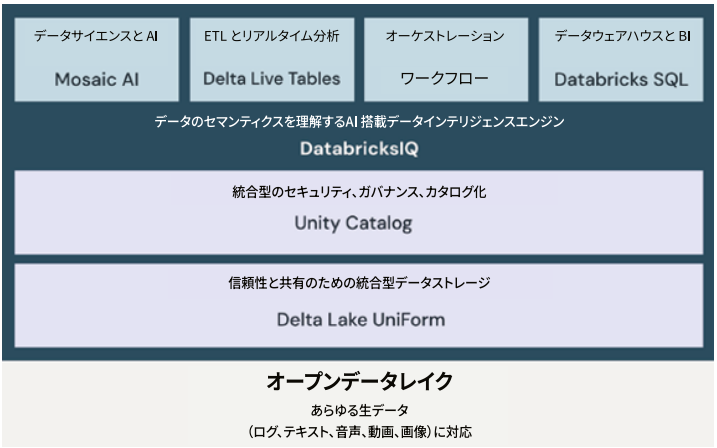


図 3-1: **Databricks** データ・インテリジェンス・プラットフォーム。

オープンデータレイク

Databricks では、独自の形式や閉鎖的なエコシステムに縛られることなく、データを常に管理下に置くことができます。データレイクは、画像、動画、音声、半構造化データ、テキストなど、新しいデータアプリケーションの多くに求められるデータタイプの保存、精製、分析、アクセスに使用できます。

Delta Lake UniForm

Delta Lake Universal Format（UniForm）を使用すると、任意の Iceberg や Hudi クライアントを使用して UC エンドポイントから Delta テーブルを読み取ることができます。DatabricksIQ は、データストレージでよくある課題の解決に AI モデルを利用するため、時間の経過とともにテーブルに変更があっても管理を手動で行う必要がなく、高速なパフォーマンスが実現します。

ストレージに使われる主な形式には、Delta Lake、Apache Iceberg、Apache Hudi の 3 種類があります。以前は、企業がデータの複製をいくつかの場所と形式で複数作成しておくのが一般的でしたが、この方法ではコストも時間もかかり、必要なコストと労力が倍増してしまうことになります。

Databricks が発表した Delta Lake UniForm では、データを 3 つの形式いずれに保存した場合でも、コピーを何度も行うことなく（ビジネスインテリジェンスや AI 用などに）データを処理できます。

Unity Catalog

Databricks データ・インテリジェンス・プラットフォーム内で、データと AI に対し統合されたガバナンスレイヤーを提供するのが Databricks UC です。UC を使用することで、組織は構造化データや非構造化データ、ML モデル、ノートブック、ダッシュボード、ファイルを、クラウドやプラットフォームを問わずにシームレスに管理できます。また、データサイエンティスト、アナリスト、エンジニアは、信頼できるデータと AI 資産を安全に発見、アクセス、コラボレーションし、AI を活用して生産性を高め、レイクハウスアーキテクチャの潜在能力を最大限に引き出すことができます。この統一されたガバナンスアプローチは、データと AI のイニシアチブを加速すると同時に、規制コンプライアンスを簡素化します。

DatabricksIQ

DatabricksIQ は UC を基盤に構築され、UC により管理されます。UC にあるデータ資産すべての説明とタグを DatabricksIQ が自動で挿入するため、UC のガバナンスが改善されます。この説明とタグが付けられたデータ資産を活用することで、専門用語、略語、メトリクス、セマンティクスがプラットフォーム全体で認識され、セマンティック検索の精度、AI アシスタントの品質、ガバナンスの実効性の向上が期待できます。

DatabricksIQ には、Databricks の製品内検索機能を大幅に強化するという効果もあります。この新しい検索エンジンは、単なるデータ検索にとどまらず、データを解釈・整列させ、実用的で文脈に沿った形式で表示するため、ユーザーがデータ作業に着手するスピードが上がります。


ヒント

UC に登録されたデータ資産は、DatabricksIQ によってユーザーが使う自然言語（NL）や企業特有の用語でデータ検索ができるよう整えられ、データの検索性が大きく改善するため、組織にあるデータ資産の利便性が向上します。

Mosaic AI

Databricks による MosaicML 社の買収により、データ・インテリジェンス・プラットフォームに同社のテクノロジーが加わり、LLM 関連機能が大幅に強化されたことで、特定のニーズに即した生成 AI アプリケーションのファインチューニングやカスタマイズを行っていただけます。この統合により、ユーザーは独自のデータのプライバシーを確保し、管理を行いながら、モデルをゼロから構築することも、既存モデルを改良することも可能となりました。

プラットフォームが生成 AI を活用することでデータの理解が深まり、インテリジェント検索機能の強化、SQL コードの作成と修正の支援、データテーブルとカラムの詳細説明の自動生成を行うセマンティックの理解が可能になります。Mosaic の特長と Databricks の生成 AI 機能を融合することで、データセキュリティとユーザーの自律性を重視した AI アプリケーション開発に最適な強力な環境が実現します。

Mosaic AI と、このプラットフォームを使った独自の生成 AI アプリケーション構築については、第 4 章で詳しく解説します。

Delta Live Tables

Delta Live Tables（DLT）は、Databricks データ・インテリジェンス・プラットフォーム向けの宣言型 ETL フレームワークです。データチームが行うストリーミングおよびバッチ ETL をシンプルにし、コスト効率を高めます。データに対して実行する変換を定義するだけで、DLT パイプラインがタスクのオーケストレーション、クラスタ管理、モニタリング、データ品質、エラー処理を自動的に管理します。

DLT では、ETL が何をすべきかを記述すれば、データインテリジェンスエンジンがデータと変換を理解し、処理のワークロードを自動でスケーリングします。DatabricksIQ がすべてを処理し、総所有コストの最適化に必要なものだけを更新します。さらに、新たなデータが追加された場合は、エンジンが基盤となるテーブルの更新に最適な方法を判断するため、ストリーミング/リアルタイム ETL を低コストで実現できます。ダウンストリームの業務アプリを有効化するには、組み込みのデータ品質とモニタリングが欠かせません。

データエンジニアがさまざまなソースからデータを抽出する際に利用するプロセスが ETL です。続いて、抽出したデータを、利用が可能な信頼性の高いリソースに変換します。最後に、そのデータをエンドユーザーがダウンストリームでアクセスして使うシステムにロードし、業務上の問題を解決するという流れになります。

Databricks Workflows

Databricks データ・インテリジェンス・プラットフォーム上でデータ処理、ML、分析パイプラインをオーケストレーションするのが Databricks Workflows です。多様なタスクをサポートしており、高度なオブザーバビリティ機能と高い信頼性を備えているため、データチームはサーバーレスコンピューティングであらゆるパイプラインの自動化とオーケストレーションを促進できます。

データインテリジェンスをコアに据えた Databricks Workflows は、見込みのある解決策を提案して、デバッグとアラートの作業をシンプルにするだけではありません。あらゆるやりとりを分析して、どのジョブとチームがデータの処理を担当しているのかを特定できるため、データ処理とオブザーバビリティも簡素化できます。ジョブが失敗した場合は、インテリジェントにタスクを復元させて必要な部分だけを再実行するため、総所有コストの大幅な削減とインテリジェンスの強化につながります。

Databricks SQL

Databricks SQL は、サーバーレスデータウェアハウスの代表的な製品です。以下はその機能の一部です。

- ETL ワークロードと BI を実行、UC によるガバナンス上のメリットをもたらす
- オープンソースの基盤アーキテクチャを使い、最適な価格とパフォーマンスで拡張
- クエリの実行速度を最適化、データ分析を容易にする
- クエリやレポートを作成するツールなど、高度な技術でデータアクセスを高速化

Databricks SQL には、Photon という名の次世代ベクトル化クエリエンジンが使われており、数千もの最適化機能が搭載されているため、あらゆるツール、クエリタイプ、現実世界のアプリケーションで最高のパフォーマンスを実現します。この最適化機能には、ニューラルネットワークを活用したデータのインテリジェントなプリフェッチにより、インデックス作成などのパフォーマンスチューニングが不要になる AI 搭載の予測 I/O も含まれます。

SQL は、その汎用性、効率性、利用範囲の広さから、データ分析に不可欠なものとなっています。また、SQL はシンプルなため、大規模なデータセットの迅速な取得、操作、管理が可能です。SQL にAI 関数を組み込んでデータ分析を行うことで、効率が高まり、インサイト抽出の迅速化を図ることができます。

AI Functions は、Databricks の組み込み型 SQL 関数で、SQL から直接 LLM にアクセスできます。AI Functions により、LLM を呼び出す際の技術的な複雑さが抽象化され、アナリストやデータサイエンティストが、基盤のインフラストラクチャを気にすることなく LLM の利用を開始できます。

DATABRICKS AI/BI について

Databricks AI/BI は、組織内のすべての人に対して BI を民主化するために、データインテリジェンスに基づいて構築された、新しいタイプのデータインテリジェンス製品です。Databricks のデータインテリジェンスエンジン DatabricksIQ を搭載した AI/BI は、お客さま固有のデータとビジネスコンセプトを理解します。これを実現するのは、データプラットフォームからシグナルを自動で取得し、精査された指示とともに明確化をプロアクティブに求め、組み込む機能です。これにより、実世界の複雑なデータから、AI が生成した関連性が高く正確なインサイトを得ることができます。

ダッシュボードでは、アナリストが自然言語を使用して、ビジネスチーム向けに高度にインタラクティブなデータ可視化を迅速に構築できます。Genie では、ビジネスユーザー自身がデータとの対話により分析を行うことができます。経験豊富な同僚に尋ねるような感覚で質問することができ、テクノロジーの専門職に頼ることなく、データから直接、信頼性の高い回答が得られます。

Databricks AI/BI は、Databricks データ・インテリジェンス・プラットフォームにネイティブに統合されており、大規模なデータでもインタラクティブなパフォーマンスを損なうことなく、即座にインサイトを提供しつつ、Unity Catalog を使用して統一されたガバナンスときめ細やかなセキュリティを確保します。

ベンダーが自社の独自システムを用いている場合、ベンダーロックインにつながる恐れがありますが、オープンソーステクノロジーを構築基盤とする Databricks は、こうした競合他社と一線を画します。このオープンなアプローチが、オープンソースコミュニティからのコントリビューションによるイノベーションを促進します。

DatabricksIQ によるプログラマー支援

Databricks Assistant は、Databricks ノートブック、SQL エディター、ファイルエディターでネイティブに利用できるコンテキスト認識 AI アシスタントです。Databricks Assistant を使えば、会話形式のインターフェースでデータを照会することができ、Databricks 内での生産性が向上します。タスクを英語で記述すると、Databricks Assistant が SQL クエリを生成し、複雑なコードを説明し、エラーを自動的に修正します。また、Databricks Assistant は UC のメタデータを活用して、テーブル、カラム、説明、および会社全体で注目されるデータ資産を理解し、あなたにパーソナライズされた回答を提供します。

SQL、Python、R、Scala のコード生成

生成 AI は、コードの生成を支援し、データクエリのプロセスを効率化します。また、データのセマンティクスとユーザーのクエリの意図を理解することで、コードの自動生成が可能で、データ操作に要する時間と労力が削減されます。

例えば、自転車販売台数トップ 10 の自治体を調べる SQL プログラムや、従業員の年間給与を隔週給与に分割する Python のプログラムを記述できます。

コードを異なる言語に変換

生成 AI の機能のひとつに、あるプログラミング言語から別のプログラミング言語へのコード変換がありますが、これは、プログラミング言語が複数使用されている環境で役立ちます。この機能により、システムとアプリケーション間のシームレスな統合と相互運用が可能になります。

既存コードの文書化や説明

プロジェクトが複雑な場合は特に、既存のコードを理解するのが大変な作業になります。コードの文書化や説明で役立つのが生成 AI で、特定のコードセグメントで何を行うのかを明確かつ簡潔に説明できます。この機能は、新しいチームメンバーのオンボーディングだけでなく、コードベースの保守と更新にも有用です。

問題やエラーのデバッグと修正

生成 AI は、コード内の潜在的な問題やエラーを特定し、デバッグと修正に関する提案を行います。エラーの検出と解決を図るこの手法で、開発時間を短縮し、ソフトウェアの品質を向上させることができます。プロンプトに「/fix」と入力するだけでコードが修正され、便利なドキュメントや詳細情報を参照できるリンクも表示されます。

文脈に即した回答が得られる

生成 AI を活用すれば、そこから得られる回答が、開発者にとって変革をもたらすツールになり得ます。生成 AI が、コーディング時のユーザー個々の傾向、プロジェクトの特性、データの意味に適応することで、あらゆるアドバイスとサポートを確実に関連付け、業務とその最新データに直接適用することが可能になります。これにより、開発プロセスが簡素化され、直観性と関連性が向上します。

本章の内容

- » 従来型 AI を使ったアプリケーション構築
- » 従来型 AI 開発にある課題に対処する
- » モデル管理に着手する
- » 生成 AI でのアプリケーション構築
- » あらゆる機能の集約

第 4 章
データ・インテリジェンス・プラットフォームでの AI アプリケーション構築

ビジネスで利用されるテクノロジーのほぼあらゆる方面で、従来型の人工知能（AI）や生成 AI が組み込まれるようになり、企業が顧客へ提供するサービスを向上させ、競争で優位に立つには、自社独自の AI アプリケーション開発が不可欠です。

本章では、Databricks のデータ・インテリジェンス・プラットフォームがどのように従来型・生成型の AI アプリケーション開発をサポートし、機械学習運用（MLOps）と大規模言語モデル運用（LLMOps）を通じてその管理を行うのかについて取り上げます。また、機能エンジニアリング、モデルの作成、モデルの実験追跡、ML の自動化、AI アプリケーションのデプロイをシームレスに実現するプラットフォームのツールと機能を掘り下げます。Databricks データ・インテリジェンス・プラットフォームは、予測モデルの構築から最新の生成 AI や LLM に至るまで、AI と ML を活用したソリューションを構築、デプロイ、監視する統合型のツールです。

従来型 AI アプリケーションの開発

従来型の AI では、明示的なプログラミングアルゴリズムを基に構築されたモデルが使われます。この形式の AI は、論理規則や意思決定のプロセスを人間のガイダンスに依存し、予測や分類といった特定の構造化されたタスクに最適化されたものが従来型の AI モデルです。データ・インテリジェンス・プラットフォームが持つ強みのひとつは、MLOps と LLMOps で、モデルの開発から、実験の追跡、モデルの管理、モデルのデプロイ、そして基盤データの変更に伴う AI モデルすべての健全性監視など、AI モデルのライフサイクル全体で力を発揮します。

Delta Live Tables と Databricks Workflows を活用する

従来型 AI アプリケーションは、Delta Live Tables と Databricks Workflows を活用して開発・デプロイの強化を図ることができます。

- **Delta Live Tables:** 信頼性の高い大規模なデータパイプラインの構築と保守をシンプル化する機能で、抽出、変換、ロード（ETL）プロセスのさまざまな作業を自動化し、データの完全性を確保するとともに、手作業による監視の必要性を減らします。Delta Live Tables では、適切な順序でオーケストレーションするタスクを定義し、データ品質に適用するルールを設定した上で、データを取り込んだ際にデータの問題を処理するといったことが可能で、何らかの原因で計画に支障が発生した場合に、根本原因をすばやく特定する堅牢なエラー処理機能を備えており、イベントログでパイプライン全体を監視する上でも役立ちます。使いやすく拡張性があり、分散環境でも高いデータ品質を確保できます。
- **Databricks Workflows:** 基盤となるインフラストラクチャの事前設定および管理が不要なジョブの実行を可能にする機能で、ワークロードに合わせリソースを自動で最適化・拡張、環境をほぼ瞬時に立ち上げることができるため、データ処理や分析パイプラインの導入が容易になります。また、コストの削減、高パフォーマンスの維持にも貢献します。

ガバナンス、セキュリティ、コンプライアンスを確保する

企業を経営する上で、AI アプリケーションの重要性が高まっており、強力なガバナンス、セキュリティ、コンプライアンスを確保する手段に対するニーズも高まっています。Databricks では、Unity Catalog（UC）を使用して、包括的なガバナンスとセキュリティ機能を提供し、データのプライバシーと規制コンプライアンスを確保できます。

これらの機能は、従来型、生成型のいずれであっても、AI アプリケーションで機密情報や社外秘の情報を扱う際に不可欠なもので、生データから AI モデル、ノートブック、アプリケーションに至るまで、すべてが責任の下に使用され、不正アクセスに対する保護が講じられていることを担保します。

従来型 AI 開発の課題に取り組む

従来型 AI アプリケーションの開発には、特有の課題が伴います。

- **データの品質と可用性の低さ:** AI モデルの基盤となるのはデータです。データが低品質で不十分な場合、AI モデルのパフォーマンス、精度、信頼性が損なわれかねません。
- **モデルの複雑さ:** 従来型の AI モデルは複雑になることがあり、理解、信頼、管理、拡張する上での難しさがあります。
- **広範なエコシステムとの統合:** AI アプリケーションを複数の社内外の業務システムやワークフローと統合させると、カスタマイズ性や設定の選択肢が広がります。

こうした課題に対処すべく、Databricks では、AI の開発ライフサイクルを合理化するよう設計された一連の機能を用いています。

- **データの管理:** Databricks は、データの統合、処理、品質管理用の包括的なツールで、AI モデルによる高品質なデータへのアクセスを可能にします。
- **AI ワークフローの簡素化:** Databricks の統合型アプローチが、複雑な AI モデルとワークフローの管理を簡素化します。生成 AI には、データの準備や初期データ分析といったタスクの自動化が可能で、こうしたタスクが利用しやすくなります。

» **シームレスな統合:** AI アプリケーションと既存のシステムの統合を容易にするさまざまなアプリケーションプログラミングインターフェース（API）とコネクタを提供し、スムーズなデプロイを実現します。

モデル管理と MLOps/LLMOps の考察

AI モデルは突然現れるわけではありません。モデルの構築、デプロイ、管理には、エンドツーエンドのプロセスがあり、そのあらゆる段階を Databricks のデータ・インテリジェンス・プラットフォームが支援します。このセクションでは、MLOps/LLOps の世界を掘り下げます。

フィーチャーエンジニアリングの改良

Databricks データ・インテリジェンス・プラットフォームの中で直接構築、あるいはサードパーティ環境への接続を経由して構築のいずれの場合でも、モデルは、レイクハウス環境にあるデータを基に構築されます。Databricks データ・インテリジェンス・プラットフォームは、データの品質を担保し、高品質のモデルを実現します。データが変換され、レイクハウスに読み込まれれば、モデルの作成に着手できます。

モデルは、既存機能の処理に基づいて新たな機能（別名：変数）を作成すると、改良や改善できることがよくあります。例えば、財務では、株価と収益の比率を基準に、株価収益を計算する機能を新規に作成することができます。そこから算出された比率には、元の株価や収益特性よりも高い予測力を期待できます。

モデルの開発

データが揃ったら、モデルの開発に着手できます。MLflow は、優れたオープンソースのモデル開発ソリューションで、Databricks にはこの機能が付属しています。特定のデータで使用するモデルの種類を選択することができ、例えば、ある製品の最適価格を予測するには線形回帰のようなモデルが適していますが、ローン希望者が融資を受ける際に所定の金利を適用すべきかを判断する場合は、ロジスティック回帰のようなモデルが適しています。

ヒント

ハイパーパラメータとは、各モデルをどのように実行するべきかの指示です。モデルに特定の振る舞いをさせるために設けておくダイヤルやレバーのようなものだとお考え下さい。例えば、顧客を分類するセグメント数を、8、20 のどちらに設定すればよいか検討しているとします。

モデルの種類とハイパーパラメータを選択すると、MLflow がモデルを実行し、与えられたデータに対するモデルの予測精度を、80%、99%といった具体的な数値で示してくれます。

実験追跡の記録

モデルを開発したからといって、それで終わりではありません。十分な実験回数で、さまざまな機能やハイパーパラメータを試しながら、どのアプローチが最も高い精度かの見極めが必要です。MLflow は、モデルが作成された時点で、その精度に関するさまざまなメトリクスを提示できます。

ポイント

モデルが満足のいくものになれば、MLflow のトラッキングサーバー経由で UC に登録します。サーバーには、モデルの作成日、バージョン番号、使用したハイパーパラメータ、精度メトリクスなどの情報が保存されます。

実験を重ね、他より優れているのはどのチャレンジャーモデル（新モデル）かを判断し、最終的にチャンピオンモデル（最良モデル）を置き換えることができます。十分な権限があれば、この比較を実施した上で、最良のモデルを採用できるということです。

AutoML による効率化

スケーリングには自動化が欠かせませんが、自動機械学習（AutoML）には、さまざまなタイプのモデルやハイパーパラメータの組み合わせを試しながら、多数のモデルシナリオを実行することが可能です。数百、数千もの組み合わせが試され、すべて自動で行われます。最後にリーダーボードが表示され、AutoML はここで与えられたデータと目的に対して最も精度が高いモデルを判断します。Databricks の AutoML により、データサイエンティストは時間のかかる反復的なタスクから解放され、複雑な作業に集中して取り組むことができます。AutoML により、これまでは専門家である博士号を持つプログラマーにしか扱うことのできなかったモデルを、テクノロジーにあまり詳しくないビジネスアナリストでも作成できるようになります。

モデルの説明可能性と透明性を明確にする

現実世界では、モデルの意思決定の要因を把握せず盲目的にモデルからの推奨事項を実施すべきではありません。ビジネスでは、成果を上げるために最も重要なのはどの機能であるか把握しておくことが求められます。銀行融資の例で考えてみます。ある見込み客に特定の金利で融資を行う場合、推奨の理由、あるいは反対する理由は何なのでしょう？

これを説明できなければ、意思決定がどうなされたのかがブラックボックスのようになり、モデルに対する人間の信頼は、当然限定的なものになります。こうしたことが原因で、モデルの実用化が見送られることは珍しくありません。

Databricks では、リネージ追跡により、モデルに使われたデータから、それが出した最終結果まで透明性が徹底化されているため、データジャーニーにおけるステップのすべてに透明性がもたらされます。

モデルをデプロイする

モデルが完成すれば、データサイエンティストや ML エンジニアは Databricks を使い簡単に本番環境へそれをデプロイできます。これは、モデルサービングエンドポイントを作成して行いますが、それほど経験を要する作業ではありません。

ML ワークフローで、モデルを実験と開発からステージング、そして最終的には実世界での使用に向けて本番環境へと移行させます。

モデルガバナンスを観測する

時が経つにつれ、最初のモデルから 2 番目、50 番目、そしてほどなく数百、数千というモデルが稼働するようになるでしょう。重要なのは、モデルのライフサイクルを管理できる状態にしておくことで、これを容易にするのが Databricks です。

また、UC によるモデルのライフサイクル管理も大切です。モデルを本番環境に配置できるのは然るべきユーザーに限られ、モデルの構築元となるデータにアクセスできるのは許可されたユーザーだけに制限されているか、確認しておく必要があります。Unity Catalog の監査ログとシステムテーブルには、誰がどのモデルをどのデータを使っていつ実行したかなど、詳細がすべて表示されます。

モデルとデータのドリフトを監視する

データは大抵、時間の経過とともに変化します。金利は変動し、購買パターンは変化し、顧客の支出も上下します。これはデータドリフトと呼ばれ、データの変化幅が大きくなると、モデルの精度の低下や価値の喪失といったことが起こり得ます。

この対策には、Databricks SQL ダッシュボードで、モデルがそれぞれ完全に健全性を維持しているか、不具合（データソースの動作停止など）を起こしていないか、モデルの更新が必要か、などを確認できます。メトリクスは開発中に設定でき、レイクハウスモニタリングのオプションでは、カスタムのメトリクスも作成可能です。


ヒント

モデルが旧式化した場合は、Databricks から SQL アラートが送信され、モデルの再調整が必要になる時期をデータサイエンティストに通知したり、人の手を介さず自動でモデル更新を行うよう設定したり、などが可能です。

生成 AI アプリケーションの開発

あらゆる業界で、企業による新たなアプリケーション開発の手法に革新をもたらしているのが生成 AI です。このテクノロジーにより、イノベーションのスピードアップ、製品のカスタマイズ、複雑な課題への取り組みが可能になります。生成 AI の採用により、企業は開発期間の短縮、ソリューションの拡張性向上を図ることができ、優れたサービスや製品の提供につながります。

カスタムの生成 AI アプリケーションを構築する

Databricks エコシステムの一端を担う Mosaic AI では、ゼロから生成 AI アプリケーションを構築できます。生データから始め、業務のコンテキストや独自のデータに対応するよう特別設計した AI モデルを開発でき、機密情報がデータ境界の外に漏洩することはありません。

Databricks データ・インテリジェンス・プラットフォームにある Mosaic AI には、主に、以下のような機能が搭載されています。

- **LLM トレーニングのカスタマイズ:** 組織独自のデータを使った LLM のカスタマイズが可能です。これにより、モデルの知識を特定のドメインに密接化し、関連性と精度の高いアウトプットが得られます。

LLM は、ビッグデータを活用して、人間が使うようなテキストを理解、生成します。言語処理を行うために膨大な情報から学習し、トレーニング用のデータから認識したパターンを基にテキストを作成します。

- **トレーニングコストの削減:** このプラットフォームは、カスタム LLM のトレーニングコストを大幅に削減する最適化済みのトレーニングソリューションを備えているため、モデルの品質を損なうことなくカスタムの AI ソリューションに投資することが、より多くの企業で可能になります。
- **包括的なモデルのサポート:** AI モデルのトレーニング後は、そのデプロイ、管理、クエリを実行するための統合型サービスを提供します。カスタム ML と基盤モデルが含まれ、ビジネスアプリケーションとワークフローへのシームレスな統合を確実なものにします。
- **データセキュリティとガバナンスの強化:** データと知的財産のすべてが自社の管理下にあることを保証し、データのプライバシーとコンプライアンスに関するリスクを低減します。医療機関や金融機関など、機密情報を扱う企業では、特に重要な機能です。さらに、データから本番環境モデルに至るまでの強力なアクセス制御、エンドツーエンドのリネージ、監査といった機能を利用できます。
- **完全な制御:** モデルとデータ両方のオーナーシップを管理します。Databricks では、組織による独自のエンタープライズデータを使った生成 AI ソリューションの構築が可能です。
- **さまざまな 生成 AI アーキテクチャのパターンをサポート:** Databricks では、プロンプトエンジニアリング、検索拡張生成（RAG）、ファインチューニング、カスタム LLM の事前学習など、複数の生成 AI アーキテクチャをサポートします。この柔軟性により、特定のユースケースに最適なアプローチを選択し、要件の変更に即した開発が可能です。

RAG アプリケーションを設計する

RAG アプリケーションは、LLM とカスタムエンタープライズデータを組み合わせ、AI が生成する回答の精度と関連性を向上させます。また、クエリに関連するデータを取得、LLM にコンテキストとして提供します。


ヒント

最新情報の管理、ドメイン固有のナレッジへアクセスを要するチャットボットや Q&A システムのサポートで力を発揮しているのが RAG で、モデル全体をファインチューニングして言語モデルをドメイン特化型アプリケーションに適応させるといった他の方法に比べ、高い費用対効果と効率が得られます。組織が基盤の LLM モデルを変更することなく外部データを利用できるため、データの頻繁な更新を要する場合に特に役立ちます。また、モデルの回答は古い可能性があるトレーニングデータではなく、最新の情報が基になっていることを担保します。

既存モデルをファインチューニングする

オープンソースの LLM モデルをすでにお持ちであれば、Databricks Mosaic AI を使い自社データで LLM モデルをファインチューニングし、事前学習済みの生成 AI モデルを特定のデータセットやドメインに適合させることができます。つまり、データセットをより適切に反映するようモデルを調整して、モデルのパフォーマンスを改善できるということです。ファインチューニングを行うことで、データに対する制御を維持でき、また、データが安全な環境から外へ出ることがないため、プライバシーを確保できます。

ゼロからモデルを構築する

Databricks Mosaic AI では、カスタムの LLM モデルをゼロから構築して、自社データの独自特性に即した AI ソリューションを作成できます。このプロセスでは、独自のデータセットで新たなモデル全体の訓練を行うため、そこから得た AI アプリケーションが自社のビジネスプロセスに統合され、確かなインサイトが生成されます。

LLM の訓練は一般に、複雑で難しく、広範な専門知識を必要とするものですが、Mosaic AI の Foundation Model Training では、誰でも簡単かつ効率よく独自のカスタム LLM をトレーニングでき、ユーザーが行うのはデータソースの指定だけです。Foundation Model Training は、数百の GPU へのスケーリング、監視、自動復旧など、残りの部分を処理します。パラメータが数十億にも上る LLM のトレーニングが、数週間ではなくわずか数日で完了します。

可能性の一例として、Databricks では過去に、Foundation Model Training と Mosaic AI のパワーを活用し、DBRX と呼ばれる最先端 LLM のトレーニングを実施しています。Mixture of Experts（MOE）アーキテクチャで構築された DBRX は、発表当時、品質と価格性能比においてトップクラスのオープンソース LLM に位置づけられるものでした。DBRX が備える技術と最適化機能のすべて、そして Foundation Model Training により、あらゆる組織が、自社のデータに合わせてフルカスタマイズした独自の LLM を、手頃な価格で構築できます。そこから、組織の知的財産を基に訓練され、独自に差別化されたカスタマイズモデルが完成しています。

すべてを統合する

Databricks データ・インテリジェンス・プラットフォームを活用すれば、エンタープライズ AI アプリケーション開発を劇的に簡素化できます。DatabricksIQ は AI プラットフォームである Mosaic AI に直接統合されているため、企業が、自社のデータを理解する AI アプリケーションを簡単に構築できます。エンタープライズデータを AI システムに直接統合させる際、Mosaic AI では以下の機能をご利用いただけます。

- カスタムデータを基に高品質の会話型エージェントを構築するエンドツーエンドの RAG
- 組織のデータでカスタムモデルをゼロから訓練する、あるいは、DBRX、MPT、Llama 3 といった既存のモデルを継続的に事前学習させ、対象ドメインの理解を深めることで AI アプリケーションのさらなる強化を図る機能
- エンタープライズデータを対象にした効率的かつセキュアなサーバーレス推論と、UC のガバナンス・品質モニタリング機能への接続
- 人気の高い MLflow オープンソースプロジェクトを基にしたエンドツーエンドの MLOps。生成されたモデルとデータはすべて、レイクハウス内で自動的に実行、追跡、監視が可能

本章の内容

- 統合型データプラットフォームのメリット
- データからインサイトを見つけ出す
- データと知的財産の所有
- コストの削減

第 5 章
データ・インテリジェンス・プラットフォームが求められる 10 の理由

企業は日々、多様なソースから大量のデータを収集していますが、膨大な量のデータにアクセスできるだけでは不十分です。データ資産が秘める能力を最大限活用するには、強力なツールが必要です。今こそ求められているのが、データ・インテリジェンス・プラットフォームです。その理由を挙げます。


ポイント

- **統合型のプラットフォームを備える:** 統合型プラットフォームが中心的役割を果たし、あらゆる種類のデータを一元管理する場所となります。一貫したデータ管理が実現し、データのサイロ化が解消されます。
- **データと AI のセキュリティが強化される:** データ・インテリジェンス・プラットフォームの利用により、データと人工知能（AI）のセキュリティが強化されます。プラットフォームに備わった強固なセキュリティ機能が、組織が機密データを保護、コンプライアンス要件を満たす上での支援となります。

  AI や 各種の分析を業務に使う企業では、情報漏洩の防止が必須です。
- **データと知的財産（IP）をすべて所有する:** プラットフォームを統合することで、自社独自のデータを基にしたアプリケーション、ソリューション、分析機能の構築や強化が可能になり、競争上、財務上の優位性が最大限強化されます。

- **検索性能が向上する:** データ・インテリジェンス・プラットフォームにより、データ資産の検索が容易になり、データの意味を適切に把握する上で必要なコンテキストを見つけ出すことができます。これが実現すれば、データインサイトが分かりやすいものになり、組織内のさまざまなユーザーによるアクセスが可能になります。

- **インテリジェントかつデータドリブンなインサイトが得られる:** データ・インテリジェンス・プラットフォームにより、データに隠されたインサイトや傾向を発見できます。データ品質の確保には、データのクレンジング、検証、エンリッチメント機能が使われるため、AI がコンテキストで認識された情報の質を高めることは重要です。

- **自動化済みの統合ワークフローでデータ関連作業がスピードアップする:** 単一のプラットフォームの中で作業を行うため、データエンジニアリング、データサイエンス、機械学習などのタスクが高速化され、シームレスなコラボレーションと効率的なワークフローが実現します。データ・インテリジェンス・プラットフォームがデータパイプラインとインフラ管理を自動化し、手作業を減らし、エラーを最小限に抑え、スケーラビリティを向上させます。

- **組織のユーザー誰もがデータにアクセスしやすくなる:** データインテリジェンスにより、ソフトウェアコードの書き方を知らない非技術系ユーザーでも、自然言語で自身の業務に関係するデータを照会できるようになり、ビジネスアナリスト、経営幹部、事業部門責任者など、誰もが以前より容易かつ迅速に、インテリジェントなインサイトを得られる可能性が開かれます。

- **効果的なコラボレーションが実現する:** プラットフォームにより、チームやユーザー間の共同作業が容易になり、インサイト、コード、分析結果の共有が活発化します。ここから生まれるチームワークが、データドリブンな意思決定を加速、ビジネスに大きなメリットをもたらします。

- **拡張性が高まる:** 大規模なデータ処理に対応するプラットフォームにより、膨大なデータの処理が効率化します。単一のプラットフォームで、構造化データ、半構造化データ、非構造化データのすべてを取得できます。

- **ROI が改善する:** コストの削減を望まない人はいません。データ管理と分析ツールを単一のプラットフォームに統合することで、コストの削減、データインフラストラクチャの簡素化を図ることができます。

データと AI で自社の可能性を最大限引き出す

データインテリジェンスは、組織のデータと AI が秘める可能性を最大限に引き出すことができます。オープンで統一され、ガバナンス機能を備えるレイクハウスアーキテクチャを基盤とする Databricks のデータインテリジェンスプラットフォームには、データインテリジェンスエンジンが搭載され、AI を駆使して固有のエンタープライズデータを基に推論を行い、お客様がデータ資産の価値を最大限に活用いただけるようにします。ETL、データウェアハウス、BI、従来型 AI、生成 AI のいずれであれ、データドリブンの成功への過程の合理化とスピードアップに、データインテリジェンスが貢献します。

本書の内容…

- データインテリジェンスの価値
- AI の持つ力と可能性
- データ・インテリジェンス・プラットフォームの各種機能
- AI アプリケーションを構築する
- データ・インテリジェンス・プラットフォームが求められる理由

databricks

Databricks のテクニカル・エバンジェリズム部門の責任者である **Ari Kaplan** は、カリフォルニア工科大学の「Alumni of the Decade」であり、シカゴ・カブスとボルチモア・オリオールズの分析部門を創設しました。Digital Media Works の創設者の **Stephanie Diamond** は、前職で AOL のマーケティングディレクターを務めており、マーケティングおよびカスタム E ブックを多数執筆しています。

動画、ステップごとの写真によるチュートリアル、ハウツー記事、そしてご購入は、Dummies.com で！

ISBN: 978-1-394-32363-0

9 781394 323630

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.